UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

UMA ABORDAGEM ALTERNATIVA DE BEHAVIORAL SCORING USANDO MODELAGEM HÍBRIDA DE DOIS ESTÁGIOS COM REGRESSÃO LOGÍSTICA E REDES NEURAIS.

Autora: Luciane de Godói Moraes
Orientadora: Professor Dra. Lisiane Priscila Roldão Selau

Porto Alegre, Julho de 2012.
Universidade Federal do Rio Grande do Sul

Instituto de Matemática
Departamento de Estatística

UMA ABORDAGEM ALTERNATIVA DE BEHAVIORAL SCORING USANDO MODELAGEM HÍBRIDA DE DOIS ESTÁGIOS COM REGRESSÃO LOGÍSTICA E REDES NEURAIS.

Autora: Luciane de Godói Moraes

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Lisiane Priscila Roldão Selau
Marcos Roberto Eilert de Souza

Porto Alegre, Julho de 2012.

Dedico este trabalho aos meus queridos pais: Ana Maria e Orides.

"É do buscar e não do achar que nasce o que eu não conhecia."
Clarice Lispector

Agradecimentos

À Deus, que em sua força natural sempre guiou meu caminho.

Ao apoio incondicional dos meus pais (Ana Maria e Orides) e irmãos (Luciano e Fabiano) para que esta etapa iniciasse e fosse concluída.

Aos meus familiares e amigos que me incentivaram em cada momento de dificuldade e entenderam os longos momentos de ausência.

A minha vó Georgina, in memoriam, por todo o carinho e apoio aos meus sonhos.

Ao maior presente que a vida me deu, meu amor, meu companheiro, meu amigo, meu tudo - Marcelo Sartori. Por ter acompanhado com paciência todos os momentos dessa etapa da minha vida, minhas decisões, indecisões, frustrações e conquistas.

Aos amigos da CEU, que sempre ficarão no coração, pelas lições de luta, perseverança e pelos momentos de alegria.

À Universidade Federal do Rio Grande do Sul (UFRGS), pelos anos de moradia e assistência. Por proporcionar um ensino de qualidade e um engrandecimento pessoal inigualável.

Aos professores e colegas do Bacharelado em Estatística, em especial a turma 2008/01, pelo conhecimento e companheirismo.

A minha orientadora Prof. Dra. Lisiane Selau pela paciência e dedicação durante a realização deste trabalho.

Aos meus colegas de trabalho pela confiança e oportunidade de executar meus conhecimentos. Pelo fornecimento de condições para que esse trabalho fosse realizado.

Sumário

Resumo *1*

1. Introdução 1

2. Fundamentação Teórica 4

2.1 Behavioral Scoring 5

2.2 Modelagem Híbrida 6

2.3 Regressão Logística 7

2.4 Redes Neurais 9

3. Sistemática para Desenvolvimento do Modelo 12

4. Resultados 16

5. Conclusões 23

Referências Bibliográficas *25*

Abstract *27*

Este artigo será submetido à “REVISTA BRASILEIRA DE ESTATÍSTICA”

UMA ABORDAGEM ALTERNATIVA DE *BEHAVIORAL SCORING* USANDO MODELAGEM HÍBRIDA DE DOIS ESTÁGIOS COM REGRESSÃO LOGÍSTICA E REDES NEURAIS.

Luciane de Godói Moraes
Lisiane Priscila Roldão Selau
Departamento de Estatística,
Universidade Federal do Rio Grande do Sul, UFRGS,
Av. Bento Gonçalves, 9500 – Prédio 43 -111, CEP 91509-900, Porto Alegre, RS,
e-mail: lu.godoi.m@gmail.com

Resumo

Com o crescimento progressivo nos volumes de concessão de crédito no Brasil, as empresas estão buscando melhorar na assertividade da concessão e agilidade na análise do crédito, não somente para novos clientes como também para clientes antigos. Técnicas quantitativas vêm sendo difundidas para a construção de modelos de previsão de risco de crédito que, baseadas tanto em informações cadastrais, quanto no histórico de relacionamento do cliente na empresa, predizem um comportamento padrão de risco. O objetivo deste artigo é propor uma sistemática para construção de modelos de previsão de risco de crédito baseados em dados comportamentais (Behavioral Scoring), utilizando um processo de modelagem híbrida de dois estágios com regressão logística e redes neurais e avaliar seu desempenho. Todas as etapas de construção do modelo são discutidas detalhadamente, sendo abordado desde o planejamento e definições do modelo até a validação da fórmula de pontuação. O modelo foi aplicado em uma amostra de 9.070 clientes de uma instituição financeira de atuação nacional. Os resultados para esse estudo específico apontaram que o método de modelagem híbrida desenvolvido apresentou superioridade às técnicas tradicionais, ressaltando que o apoio dos resultados da regressão logística, como nós de entrada da rede neural, contornaram as características indesejáveis das redes neurais, como processamento lento e dificuldade na interpretação das variáveis.

Palavras-chave: análise de crédito, behavioral scoring, modelagem híbrida, regressão logística, redes neurais.

1. Introdução

No Brasil, a concessão de crédito sempre foi lenta e escassa, devido a políticas mal concebidas e ao processo inflacionário do passado. Entretanto, em consequência da maior estabilidade da economia brasileira nos últimos anos, após a implantação do Plano Real, as empresas têm percebido o crédito como um gerador de riquezas e de novos negócios (GOLDBERG *apud* BUENO, 2003).

Junto ao aumento da demanda pelo crédito e a grande competição industrial, surgiu a necessidade de tomar melhores decisões sobre o crédito, principalmente quanto à análise do seu risco. Jorion (1997) afirma que esse risco pode ser definido como a possibilidade de a contraparte não cumprir as obrigações monetárias contratuais relativas às transações financeiras. Esse não cumprimento das obrigações contratuais é chamado de inadimplência e deve ser monitorado.

Dessa forma, em vez de visar todos os clientes ou fornecer o mesmo incentivo a todos, as empresas podem selecionar apenas aqueles clientes que atendem a certos critérios de rentabilidade com base em suas necessidades individuais ou comportamentos de compra (DYCHE; DYCH, 2001). Para isso, é impossível, tanto em termos econômicos quanto humanos, realizar tantas análises de maneira subjetiva, principalmente pelo fato desses ambientes serem dinâmicos e com constantes alterações, em que as decisões devem ser tomadas rapidamente e em número cada vez maior (MENDES FILHO *et al.*, 1996).

Portanto, algumas autoridades bancárias, como o "*Bank of International Settlements* (BIS)", o Banco Mundial e o FMI, desenvolveram uma série de lições, todas encorajando as instituições financeiras a desenvolverem modelos internos para melhor quantificar os riscos financeiros (EMEL *et al.*, 2003). Assim, instituições financeiras de muitos países, entre eles o Brasil, estão intensificando e aperfeiçoando metodologias de estudos sobre práticas que auxiliam no controle dos riscos; tendo, como ferramentas poderosas para apoio à decisão de crédito, os métodos estatísticos e as abordagens de inteligência artificial (THOMAS, 2000).

Esses métodos são usados na elaboração dos modelos de *Credit Scoring* que consistem em uma das principais ferramentas de suporte à concessão de crédito. O desenvolvimento desses modelos, segundo Louzada (2008), baseia-se na construção de um procedimento formal para determinar quais características do cliente estão relacionadas, significativamente, com o seu risco de crédito e qual a intensidade e direção desse relacionamento. O objetivo básico é a geração de uma pontuação, pela qual os clientes possam ser classificados conforme a sua chance de inadimplência.

Modelos de *Credit Scoring* se subdividem em *Application Scoring e Behavioral Scoring. O*s modelos desenvolvidos com base nos dados de solicitação de abertura de novas contas são chamados de *Application Scoring* e quando essas contas começarem a amadurecer pode ser desenvolvido modelos de *Behavioral Scoring*

(MORRISON, 2003). Então, com a necessidade de automatizar o processo de decisão de crédito, juntamente com o aumento maciço de informações sobre as transações dos consumidores bancários (AKHAVEIN, 2005), a utilização de modelos comportamentais (*Behavioral Scoring*) é cada vez mais necessária às organizações, apesar de menos conhecidos, mas igualmente importantes, visto que até são mais eficazes do que modelos que consideram apenas dados cadastrais (*Application Scoring*) (THOMAS, 2000).

Desta forma, enquanto o principal objetivo dos modelos de aprovação de crédito é estimar a probabilidade de um novo solicitante de crédito se tornar inadimplente com a instituição em determinado período, os modelos comportamentais estimam a probabilidade de inadimplência daqueles que já são clientes, já que são incorporadas variáveis que retratam a história do cliente com a instituição, além das informações cadastrais. Assim, segundo Thomas (2000), a pontuação de crédito para já clientes ajuda tanto na decisão referente a novo limite de crédito, na identificação dos clientes que são mais rentáveis, nas ações de *marketing,* na oferta de novos produtos, como também na cobrança, na mensuração do risco e no controle das perdas; enfim, em todas as decisões relativas ao gerenciamento do crédito de clientes que já possuem uma relação ou um histórico com a instituição.

Para tanto, geralmente duas ferramentas estatísticas são mais comumente utilizadas, a análise discriminante e a regressão logística (THOMAS, 2000). Recentemente, as redes neurais estão se tornando também uma alternativa muito popular nas tarefas de análise de crédito (CORRAR, *et al*., 2007) e apresentam grandes vantagens em relação às técnicas convencionais (WEST, 2000; LEE *et al.*, 2002; BAESENS *et al.*, 2003; SELAU, 2011) como capacidade de processamento paralelo, generalização e acurácia na classificação não linear (LEE; CHEN, 2002).

Contudo, diversos pesquisadores, como Chung e Gray (1999), Hand e Henley (1997) criticam o tempo elevado de treinamento das redes neurais, limitando, com isso, a aplicabilidade na manipulação de problemas de modelagem de crédito. Dessa forma, percebe-se na literatura que, técnicas estatísticas e de aprendizagem de máquina para classificação de crédito têm sido extensivamente estudadas, sendo que recentemente os estudos estão focados em modelos híbridos, combinando diferentes técnicas de aprendizagem que têm mostrado resultados promissores. Tsai e Chen (2010) inovaram testando várias combinações de modelos híbridos, enquanto outros

pesquisadores mostraram a eficiência desses modelos em comparação às técnicas individuais (LEE, *et al.*, 2002; HSIEH, 2005; JAIN, *et al*., 2007; KIM, *et al*., 2007; CHEN, *et al*., 2009). Seu estudo indicou que a análise de regressão logística, utilizada como primeiro componente, combinada com redes neurais como o segundo componente foi superior aos outros modelos, tanto em discriminação, quanto em maximazação do lucro. Neste tipo de modelagem as variáveis significativas obtidas na regressão logística são utilizadas como nós de entrada do modelo de redes neurais, a fim de melhorar a decisão da estrutura da rede e dar suporte às dificuldades de interpretação dos resultados obtidos (LEE, *et al.*, 2002; CHEN, *et al*., 2009; TSAI, *et al*., 2010; WANG, *et al*., 2011; GHODSELAHI, 2011).

Sendo assim, o objetivo desse artigo é propor uma sistemática para construção de modelos de previsão de risco de crédito baseados em dados comportamentais, utilizando um processo de modelagem híbrida de dois estágios com regressão logística e redes neurais.

Este artigo está organizado em cinco seções. Após a introdução apresentada nesta seção inicial, a segunda seção traz a fundamentação teórica, onde é exposto o referencial sobre *Behavioral Scoring*, modelos híbridos e as técnicas utilizadas para sua construção. Na terceira seção é detalhada a sistemática proposta para a construção do modelo. Na quarta seção são apresentados os principais resultados da construção do modelo e avaliação do desempenho em um banco de dados de clientes de uma instituição financeira de atuação nacional. Na última seção são apresentadas as considerações finais do estudo, as principais conclusões obtidas e a discussão das possíveis pesquisas futuras.

2. Fundamentação Teórica

Modelos de previsão de risco de crédito vêm sendo amplamente estudados e ganhando forças devido a sua importância para a saúde de instituições financeiras, já que o sucesso dessas instituições está diretamente relacionado a sua capacidade de gerir os riscos (GHODSELAHI, 2011). Para lidar com estes desenvolvimentos, estão sendo utilizadas ferramentas matemáticas e estatísticas cada vez mais sofisticadas, sendo que, conforme Tsai e Chen (2010), uma pequena melhora na precisão da classificação de crédito pode resultar numa grande redução do risco e gerar significativa economia para a instituição.

2.1 Behavioral Scoring

Para que se possa ir além de apenas identificar os riscos dos clientes, a empresa precisa deter informações do comportamento do cliente (THOMAS *et al.*, 2001), com variáveis que demostrem o histórico com a empresa concedente, os atrasos de pagamento do cliente, a utilização média do limite de crédito, o tempo de relacionamento com a empresa, entre outras (MCNAB; WYNN, 2000; SECURATO, 2002) e/ou buscar em dados externos informações comportamentais do cliente. Percebendo esse ciclo do cliente na instituição, vários autores, como Thomas *et al.* (2001); Hand (2001); Morrison (2003); Sarlija *et al.* (2009), dividem os modelos de *Credit Scoring* em duas categorias principais: (i) *Application Scoring* – pontuação de um novo cliente, que considera variáveis cadastrais, como sexo, idade, escolaridade e indica um fenômeno estático e (ii) *Behavioral Scoring* – pontuação de um cliente antigo, que considera variáveis comportamentais, como movimentação financeira, quantidades de parcelas pagas em atraso, além das variáveis cadastrais e indica um fenômeno dinâmico. Outros autores, como Sicsú (2010) e Louzada (2008), acrescentam uma terceira categoria: *Collection Scoring* – pontuação utilizada para prever eventos futuros em relação à cobrança de inadimplentes. Assim, a principal diferença entre esses modelos está no conjunto de variáveis disponíveis para estimar a qualidade de crédito do cliente, ou seja, quanto mais precoce o estágio do ciclo de crédito, menor o número de informações específicas sobre o cliente de que dispõe a instituição.

Para construção de modelos *Behavioral Scoring* é necessário escolher um ponto de observação (THOMAS, 2000), sendo que é preciso haver dados sobre o comportamento do cliente antes e após esse ponto (THOMAS, 2000; MCNAB; WYNN, 2000; ANDERSON, 2007). O período de tempo de observação anterior ao ponto é chamado de período de desempenho ou de observação por Thomas *et al.* (2001) e é geralmente 6 a 12 meses, já Sarlija *et al.* (2009) nomeam esse período como sendo de execução, usando 6 meses em seu estudo. As características observadas durante este tempo que precede o ponto de observação serão utilizadas para o desenvolvimento do modelo. O período após o ponto de observação é o período de resultados, que normalmente é tomado como 12 meses (THOMAS *et al.,* 2001), sendo que Sarlija *et al.* (2009) usaram 6 meses, e é nesse período que o cliente é classificado como bom ou mau pagador, dependendo de seu estado no final. Thomas *et al.* (2001) alertam que uma das desvantagens para construir um

Behavioral Scoring é que normalmente é necessário histórico de dois anos e, portanto, a população na qual aplica-se o modelo pode ser bastante diferente daquela de quando foi construído. Talvez por isso, Sarlija *et al.* (2009) utilizam um período de 6 meses antes e depois do ponto de observação.

2.2 Modelagem Híbrida

Em mineração de dados, a abordagem de hibridação tem sido uma área de pesquisa ativa para melhorar a classificação, a previsão e o desempenho de modelos de pontuação de crédito. Em geral, segundo Tsai e Chen (2010), o modelo híbrido é baseado na combinação de duas técnicas diferentes, podendo ser técnicas de agrupamento ou de classificação. Análise Discriminante, Regressão Logística, Árvore de Decisão, Redes Neurais e Redes Bayesianas são exemplos de técnicas de classificação; enquanto *K-means* é exemplo de técnica de agrupamento.

Modelos híbridos vêm sendo utilizados, principalmente, para melhorar inconvenientes das técnicas de inteligência artificial (LEE *et al.*, 2002), já que a primeira técnica servirá para orientar o processamento da segunda (GHODSELAHI, 2011), diminuindo o tempo de processo e facilitando, através do primeiro método, a identificação da relevância das variáveis significativas (LEE; CHEN, 2005).

Para demonstrar a viabilidade e eficácia da proposta da modelagem híbrida em duas etapas, é apresentada na Tabela 1 uma lista de alguns artigos que usaram modelos híbridos para a pontuação de crédito, apesar de ainda haver poucos estudos enfocando esse desenvolvimento. Nesses trabalhos os modelos híbridos são comparados com alguns modelos singulares de diferentes técnicas quantitativas, sendo que todos os pesquisadores concluíram que modelos híbridos aumentam a precisão da classificação em relação a modelos de apenas uma etapa.

Percebendo essa corrente de estudos, Tsai e Chen (2010) inovaram comparando diferentes modelos híbridos para pontuação de crédito a fim de identificar a melhor combinação em termos de precisão da previsão, da taxa de erro e do lucro máximo. Verificaram em seus estudos que a regressão logística utilizada como primeira componente combinada com redes neurais como segundo componente é superior aos demais modelos. Portanto, o presente artigo apoia-se na conclusão de Tsai e Chen (2010), usando o resultado da regressão logística como nós de entrada da rede neural, garantindo uma melhor interpretação do modelo e menor tempo de

processamento, além dos benefícios já mencionados. As duas técnicas que serão utilizadas seguem descritas na sequência.

Tabela 1 - Autores que utilizaram modelagem híbrida para a pontuação de crédito.

Pesquisador	Técnicas	Avaliação
Lee *et al.* (2002)	Análise Discriminante + Redes Neurais	Precisão; Taxa de Erro; Interpretação; Convergência
Hsieh (2005)	Cluster + Redes Neurais	Precisão; Taxa de Erro
Lee e Chen (2005)	Regressão Multivariada Não-Paramétrica (MARS) + Redes Neurais	Precisão; Taxa de Erro; Interpretação
Huang *et al*. (2004) Zhang *et al.* (2008)	Algoritmos Genéticos + SVM (*Suport Vetor Machine*)	Precisão

2.3 Regressão Logística

A técnica de regressão logística surgiu apenas nos anos 80 com estudos de Ohlson (1980), justificando o uso pela imponência sobre da análise discriminante de algumas condições para as variáveis preditoras, como: normalidade na distribuição dos erros e matrizes de variância-covariância iguais entre os grupos analisados; e também pela baixa interpretação intuitiva fornecida pelo escore da análise discriminante.

Em 1989, Hosmer e Lemeshow também mostraram que a utilização da técnica de regressão logística é adequada em muitas situações, porque permite que se analise o efeito de uma ou mais variáveis preditoras (categóricas ou métricas) sobre uma variável resposta dicotômica, representando a presença (1) ou ausência (0) de uma característica. A regressão logística tem por objetivo encontrar um modelo explicativo para o comportamento da probabilidade de sucesso, em termos das variáveis preditoras. Dessa forma, a regressão logística é especificamente desenhada para prever a probabilidade de um evento ocorrer, sendo essa probabilidade classificada entre o intervalo 0 e 1.

Sendo assim, o valor esperado das variáveis preditoras passa por um processo de transformação logística em que são transformadas numa razão de probabilidades e posteriormente em uma variável de base logarítmica. Portanto, devido à natureza não linear dessa transformação, utiliza-se o método de máxima verossimilhança, em vez

de utilizar o método de mínimos quadrados utilizado na regressão linear, para estimar os coeficientes (HOSMER; LEMESHOW, 1989; HAIR *et al.,* 2005).

O modelo de previsão da regressão logística pode ser obtido pela equação 01 e 02.

$$p = 1 / (1+ e^{-Z}) \quad (01)$$

$$Z = ln\ (p / (1 - p)) = b_0 + b_1.X_1 + b_2.X_1 + \ldots + b_n.X_n \quad (02)$$

onde:

p = probabilidade do evento ocorrer;

$1 - p$ = probabilidade do evento não ocorrer;

X_i = variáveis preditoras;

b_i = coeficientes a serem estimados para cada uma das variáveis.

Notamos que, a função p normaliza a saída do modelo para o intervalo [0,1], informando a probabilidade do evento de interesse. Para testar a significância dos coeficientes, utiliza-se a estatística de Wald (HAIR *et al.*, 2005). Com o uso dessa estatística, o teste de hipóteses pode ocorrer como em regressão múltipla.

Corrar *et al*. (2007), destacam a técnica de regressão logística, pela possibilidade de contornar certas restrições encontradas em outros modelos multivariados. Contudo, o modelo de regressão logística é sensível à colinearidade entre as variáveis (HAIR *et al.,* 2005). Por isso, Corrar *et al*. (2007) indicam, como uma das ações corretivas para os problemas de multicolinearidade, o uso do método *stepwise* para escolha de variáveis que irão compor o modelo considerado. O procedimento de avaliação das variáveis preditoras desconsidera variáveis que apresentem sinais de multicolinearidade, optando por manter no modelo apenas aquelas de maior significância estatística (SELAU, 2011).

Portanto, de maneira geral, a regressão logística permite identificar e remover características que já foram detectadas por outras variáveis e assegura que toda característica importante do cliente permaneça na sua pontuação (THOMAS *et al.*, 2002).

2.4 Redes Neurais

Redes neurais é uma técnica de tratamento de dados que recentemente tem despertado interesse não apenas de pesquisadores da área de tecnologia, como também da área de negócios (CORRAR et al., 2007), devido ao seu desempenho ser muitas vezes superior a outros métodos de estatística multivariada (WEST, 2000; LEE *et al.*, 2002; BAESENS *et al.*, 2003; SELAU, 2011).

As redes neurais são sistemas de inteligência artificiais, inspirados no funcionamento de um cérebro humano, contendo propriedades particulares como capacidade de aprendizado, de generalização, ou de organização dos dados (KRÖSE; SMAGT, 1996; KHONEN,1988). Além da capacidade de aprendizado, onde os erros de saída retornam ao início da rede e são ajustados adequadamente, a rede neural tem outra vantagem sobre as técnicas estatísticas, de não necessitar de uma suposição inicial de um modelo probabilístico pré-estabelecido para estimar os parâmetros de um suposto modelo.

Segundo Gonçalves (2005), há quatro etapas para a estrutura e operação das redes neurais: (i) o processamento das informações ocorre dentro dos chamados neurônios; (ii) os estímulos são transmitidos pelos neurônios por meio de conexões; (iii) cada conexão associa-se a um peso, que numa rede neural padrão multiplica-se ao estimulo recebido; e (iv) cada neurônio contribui para a função de ativação para determinar o estimulo de saída.

Haykin (2001) apresenta as unidades básicas da rede, onde existem três camadas: a de entrada, a intermediária (onde o processo é refinado) e a de saída. Nestas camadas encontram-se os neurônios (Figura 1), que são conectados por sinapses ou pesos.

Castro Junior (2003) apresenta que para os processamentos que ocorrem em cada neurônio tem-se uma soma ponderada das entradas e calcula-se um valor de saída, resultando na equação 03.

$$v_k = \mathrm{T}\ (\sum w_{kj}\ .\ x_k\ ) \qquad (03)$$

Sendo x_k o elemento de saída; w_{kj} o coeficiente de ponderação entre os elementos *k* e *j*, que representam o conhecimento obtido pela rede; *T* é uma função de transferência também chamada de função de transformação e v_k será a função de ativação. Em

geral as funções de ativação do tipo sigmóides são as mais usadas, por serem limitadas e possuírem derivada contínua, exigência para o uso de algoritmos do tipo *Backpropagation* utilizados em múltiplas camadas (RUMELHART; McCLELLAND, 1986).

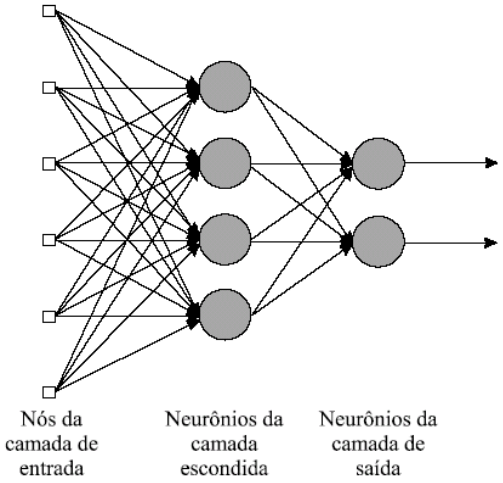

Figura 1 - Modelo estrutural de uma rede neural

Fonte: Haykin (2001).

Inicialmente a camada de entrada era ligada diretamente à camada de saída, obtendo apenas uma camada, chamadas de *Perceptron*, sendo eficazes apenas para conjuntos de treinamento linearmente separável (ROSENBLATT, 1958). Em decorrência disso, foram desenvolvidas as redes *Multilayer Perceptron (MLP),* que possuem uma ou mais camadas intermediarias (ocultas), permitindo, assim, a solução para conjuntos de dados não separáveis linearmente.

O algoritmo mais conhecido e considerado essencial para treinamento de redes neurais é o *Backpropagation* (FAUSETT, 1994; YU *et al.*, 2002), desenvolvido com o objetivo de obter os pesos que minimizem a função de erros apresentada na equação 04.

$$E = 1/2 \sum (t_j - y_i)^2, \text{ para } 1 \leq j \leq m. \quad (04)$$

Para a minimização da função calculam-se as derivadas de E com relação aos pesos e vieses da superfície de erros.

Segundo Loesch e Sari (1996), o algoritmo *Backpropagation*, pode ser dividido em 5 passos: (i) apresentação de um padrão de entrada e da saída desejada; (ii) cálculo dos valores de saída; (iii) ajuste dos pesos da camada de saída; (iv) ajuste de pesos

das camadas escondidas; e (v) verificação da magnitude do erro. Na Figura 2 temos a ilustração de uma rede *MLP* e a retro-propagação dos erros.

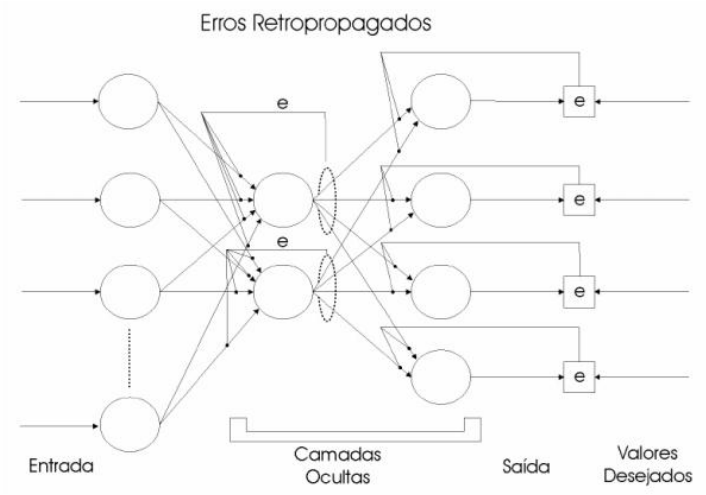

Figura 2 - Rede MLP e retro-propagação dos erros.
Fonte: Patterson (1995).

Apesar das redes neurais terem se estabelecido como uma alternativa aos tradicionais modelos estatísticos (CORRAR *et al.*, 2007; LAHSASNA, 2010) e de muitos estudos na área de crédito terem concluído que as redes neurais superam os tradicionais métodos estatísticos em termos de precisão de classificação (WEST, 2000; LEE *et al.*, 2002; BAESENS *et al.*, 2003; SELAU, 2011), Hair *et al.* (2005) alerta para a aplicação de redes neurais em problemas que necessitem previsão e classificação, com interesse na precisão da classificação e não na interpretação da variáveis preditoras. Outros pesquisadores, como Chung e Gray (1999), Hand e Henley (1997) também criticam o tempo elevado de treinamento das redes neurais, limitando, com isso, a aplicabilidade na manipulação de problemas de modelagem de crédito, apesar de que a tecnologia de processamento evoluiu significativamente, desde então.

Por isso, nesse artigo, propõe-se uma abordagem de modelos híbridos, que conforme já mostrado aqui, vêm sendo utilizada para melhorar inconvenientes das técnicas de inteligência artificial, diminuindo o tempo de processo e facilitando, através do primeiro método, a identificação da relevância das variáveis significativas.

3. Sistemática para Desenvolvimento do Modelo

Para elaboração do modelo a ser testado, optou-se por seguir as etapas de desenvolvimento de um sistema de *Behavioral Scoring* baseado em Sicsú (1998) e Sicsú (2010), que consistem em 7 etapas principais, expostas na Figura 3.

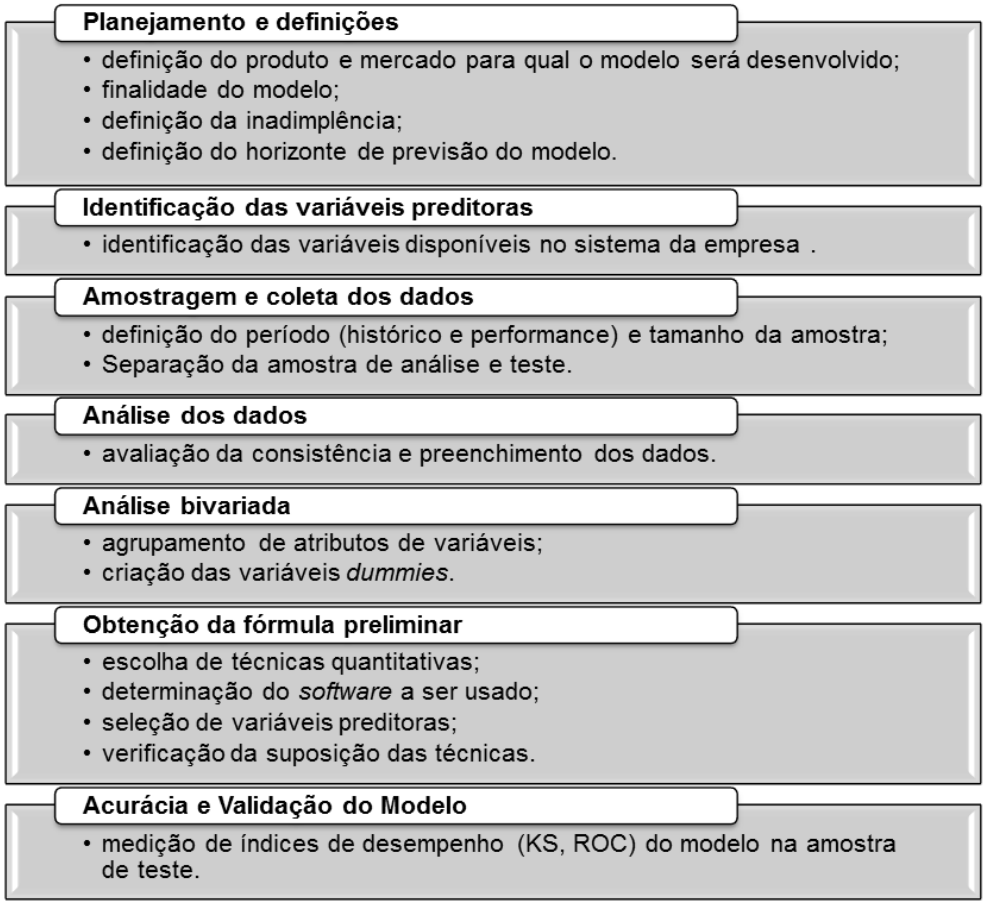

Figura 3 - Etapas para o desenvolvimento do modelo.

Planejamento e definições

Deve estar claro qual o público a ser atingido pelo modelo, ou seja, se é necessário limitar o modelo para determinado produto, o que geralmente ocorre quando há uma grande diversificação. Tendo claro o público, definir os conceitos para divisão dos grupos em termos de aceitação de desempenho. Deve haver a divisão em 3 diferentes faixas de atrasos: (i) Bons – clientes com pouco ou nenhum atraso no período de observação; (ii) Indeterminados – clientes com atrasos definidos como intermediários; (iii) Maus - clientes com atrasos significativos (SELAU,2011). A relevância dos atrasos deve seguir a regra já existente em cada instituição, conforme seu negócio e produto. Consideram-se, para a construção do modelo, apenas os

grupos de bons e maus clientes, a fim de intensificar a separação dos perfis (SELAU,2011; SICSÚ,2010).

Identificação das variáveis preditoras

Para o desenvolvimento de um *Behavioral Scoring*, além das tradicionais variáveis disponíveis nos sistemas de cadastro como: sexo, idade, endereço, estado civil, entre outras; também devem ser incluídas variáveis que traduzam o comportamento do cliente na empresa, como: tempo de relacionamento, histórico de crédito, histórico de inadimplência, histórico de investimento, entre outras. Nessas variáveis históricas, é interessante avaliar suas "derivadas", como a soma e a média das mesmas em determinados períodos (SICSÚ, 2010; THOMAS, 2000). Muitas vezes esses históricos, assim como as suas quebras, não estão consolidados nos sistemas das empresas, necessitando a elaboração de cada variável.

Amostragem e coleta dos dados

Para construção de modelos *Behavioral Scoring* é necessário escolher um ponto de observação, sendo que é preciso haver dados sobre o comportamento do cliente antes e após este ponto. O período de tempo de observação anterior ao ponto é chamado de período de desempenho histórico e é geralmente de 6 a 12 meses. As características observadas durante este tempo que precede o ponto de observação serão utilizadas para o desenvolvimento do modelo. O período após o ponto de observação é o período de resultados, que normalmente é tomado também como 6 a 12 meses, e é nesse período que o cliente é classificado como bom ou mau, conforme ilustrado na Figura 4.

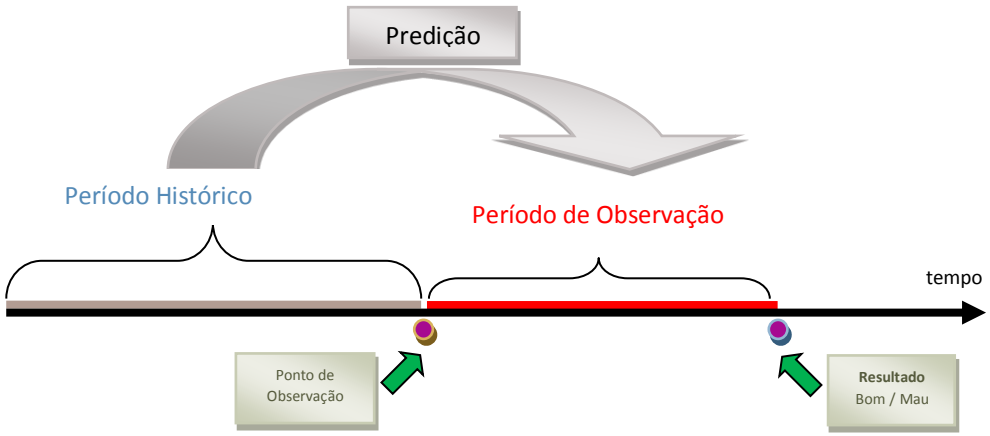

Figura 4 – Esquema para obtenção da amostra do modelo.

Repete-se essa seleção para diferentes pontos de observação, em datas diferentes, objetivando a construção de diversas safras. Muitos autores selecionam as safras a fim de controlar possíveis sazonalidades (SICSÚ, 2010), outros apenas para aumentar o volume de dados a serem estudados (SICSÚ, 2010; THOMAS *et al.*, 2001).

Tendo sido definidas a população e as variáveis, deve-se selecionar uma amostra desses clientes com um tamanho adequado para que possam ser analisados estatisticamente, geralmente utiliza-se uma proporção de 20 observações para cada variável a ser analisada (SELAU, 2011). Por fim, necessita-se do particionamento da amostra, sendo reservada uma parte dela para análise e outra para a validação/teste do modelo treinado, que será usada na verificação do poder preditivo do modelo. Sugere-se que a haja uma proporção entre 70% e 80% da amostra de análise, devido a maior importância na construção do que na validação do modelo (LOUZADA *et al.*, 2008; SICSÚ, 2010; SELAU, 2011).

Analise dos dados

A avaliação da consistência e do preenchimento dos dados é imprescindível, estatísticas descritivas podem ajudar a detectar essas inconsistências. A detecção da presença de valores discrepantes (*outliers)* é muito importante, pois mesmo sendo reais, podem comprometer a estimativa dos pesos das variáveis (SICSÚ, 2010). Além disso, devem ser avaliados os valores faltantes (*missing*), podendo efetuar exclusão de observações, ou utilizar métodos de imputação. Variáveis com muitas observações faltantes devem ser descartadas da análise. Nessa etapa também se pode elaborar novas variáveis a partir da combinação de outras já coletadas.

Analise bivariada

O objetivo aqui é confrontar cada variável explicativa com os grupos determinados de bons e maus pagadores e avaliar o poder discriminador individual. Quanto maior a diferença entre o percentual de bom e mau, maior será a contribuição da variável para a predição. O agrupamento das classes podem ser selecionado por vários critérios, entre os mais conhecidos está o *Information Value* de Kulbak (SICSÚ, 2010).

O *Information Value (IV)* utiliza-se da soma dos pesos das evidências *(WOE),* que nada mais é do que o logaritmo natural da probabilidade de ser bom dada a categoria, pela probabilidade de ser mau (equação 05), dada a categoria, ponderados após o agrupamento das categorias, conforme equação 06.

$$WOE = ln\ [\ P(c \mid Bom)\ /\ P(c \mid Mau)\ ] \quad (05)$$

$$IV = \sum [\ P(c \mid Bom) - P(c \mid Mau)\ ]\ .\ WOE \quad (06)$$

Após a avaliação dos melhores níveis de cada variável, deve-se criar a variável *dummy* para cada atributo que fará parte da análise multivariada. Essa variável assumirá apenas o valor 0 ou 1 (ex.: estado civil solteiro = 1, caso contrário =0). Dessa forma, os problemas de não linearidade serão evitados (SELAU, 2011).

Obtenção da fórmula preliminar

A determinação dos pesos de cada atributo para o cálculo do escore final é realizado com base nas técnicas quantitativas explicadas na fundamentação teórica (Seção 2). Hoje em dia há muitas técnicas para cálculo de escore, contudo, cabe ao analista verificar suas necessidades e o poder de cada uma delas. Nesse estudo sugere-se o uso de uma modelagem híbrida em duas etapas, sendo a primeira uma regressão logística, cujos resultados e variáveis selecionadas servirão de entrada para uma rede neural.

A complexidade de execução do modelo adequado deve ser sempre levada em conta, não subestimando nenhuma etapa, a fim de obter o melhor resultado. Para isso, os pressupostos de cada técnica devem ser seguidos. Na regressão logística deve-se atentar para a verificação da ausência de multicolinearidade, e para isso pode-se lançar mão do método *stepwise,* já presentes em muitos *softwares* estatísticos, assim as variáveis preditoras serão incorporadas no modelo automaticamente. Já na rede neural há uma flexibilidade, não necessitando nenhuma verificação prévia quanto a pressuposições para o uso.

A escolha do *software* também é um passo que deve ser realizado com cuidado, dado que é necessário verificar os recursos e características de cada um.

Analise da validação da fórmula de pontuação

Há muitas técnicas estatísticas para testar o desempenho dos modelos e auxiliar na escolha, entre elas: (i) estatística de Kolmogorov-Smirnov (KS); (ii) área abaixo da curva ROC (AUC); e (iii) índice de Gini.

A estatística KS é construída calculando a máxima diferença entre as distribuições acumuladas de bons e maus. A Tabela 2 indica os níveis de aceitação para um modelo construído com base no comportamento do cliente em termos de KS. Esse nível de aceitação pode variar conforme a margem de lucro da empresa, o valor

médio dos produtos, a área de aplicação do modelo, entre outros fatores; ou seja, classificação mostrada na Tabela 2 é apenas um indicativo da qualidade do modelo, cabendo ao analista avaliar a relevância do resultado em seu negócio.

Tabela 2 – Valor de KS x Capacidade de discriminação.

KS	Característica da Discriminação
KS<=40	Discriminação baixa
40<= KS<50	Discriminação aceitável
50<= KS<60	Discriminação boa
60<=KS<70	Discriminação muito boa
KS>=70	Discriminação excelente

Fonte: Sicsú (2010).

A análise da curva ROC (*receiver operating characteristic plots*) baseia-se na sensibilidade e na especificidade. Sensibilidade pode ser entendida como a capacidade de identificar os maus créditos; já a especificidade como a capacidade de identificar os bons créditos. Para um determinado escore X, a especificidade é medida pela relação de bons corretamente classificados, ou seja, a proporção de bons cujo escore é maior ou igual a X e a medida (1-especificidade), utilizada na curva ROC, representa os bons classificados como maus. Um gráfico dos resultantes pares de sensibilidade e 1-especificidade constitui uma curva ROC (LOUZADA, *et al*.; 2008). Para calcular a área abaixo da curva ROC (AUC: *Area Under Curve*), é computada a sensibilidade e a especificidade para cada valor de X, variando do menor para o maior escore. Assim, quanto maior, melhor o modelo, sendo que se o valor for acima de 0,7 o modelo é aceitável (SICSÚ, 2010).

O índice de GINI refere-se a duas vezes a área entre e a diagonal que cruza o gráfico em 45 º e a curva ROC, esse coeficiente sumariza o desempenho do modelo sobre todos os pontos de corte. Quanto maior este valor, melhor a predição do modelo.

4. Resultados

Os resultados serão apresentados na mesma sequência dos passos mostrados na sistemática da seção anterior, de forma que fiquem evidenciadas as etapas percorridas para a obtenção do modelo.

Planejamento e definições

Devido à diversidade de público atendida e da gama de produtos, procurou-se delimitar o alvo do modelo. Assim o foco foi na avaliação do risco do cliente, apenas

pessoa física, e que já tiveram alguma operação de crédito pessoal durante o tempo de análise. Dessa forma, o modelo servirá de apoio às novas decisões de concessão de crédito pessoal para já clientes, contribuindo para o controle dos atrasos de pagamentos. Conforme o conhecimento de analistas da instituição, os clientes foram divididos em três grupos, de acordo com o tempo de atraso, nos últimos 6 meses que antecederam a observação, sendo eles: (i) Bom: cliente com atraso até 30 dias; (ii) Intermediário: cliente com atraso entre 31 e 60 dias; e (iii) Mau: cliente com atraso superior a 60 dias.

Identificação das variáveis preditoras

Como o modelo a ser construído trata-se de um *Behavioral Scoring* necessita-se de informações históricas além das tradicionais do cadastro, como idade, estado civil, CEP, renda, patrimônio, entre outras. Portanto houve a necessidade da consolidação desses dados para obtenção das variáveis históricas comportamentais. Com a varredura de bases mensais houve a transformação de variáveis que consolidam o comportamento do cliente na instituição. Essa etapa também é muito importante, pois dela depende não só a construção das variáveis que resumem a totalidade da ação no tempo observado, como também suas possíveis derivações. Ao exemplo da informação de investimento do cliente ao longo do período histórico observado, pode-se apenas criar uma variável binária, indicando que o cliente teve ou não algum investimento durante o período histórico, que nesse estudo foi de 6 meses; ou ainda, sobre a mesma informação, criar outra variável que compute o total investido nesse período; ou que demonstre a média de investimento, como também poderá haver quebras em decorrência do tempo observado. Todas essas serão variáveis candidatas a entrarem no modelo, sendo que a mais significativa para discriminação será escolhida. Seguindo esses passos de decomposição das informações disponíveis, foram obtidas 82 variáveis para o inicio do estudo, sendo dessas 40 cadastrais e 42 históricas. Todas as variáveis ainda passarão pela avaliação de *outliers,* de *missings* e pela categorização (quando necessária), para então ser medida a importância de cada uma (individualmente e no conjunto) na discriminação final de bom e mau pagador.

Amostragem e coleta dos dados

Nesse estudo foi adotada a prática de observação de 6 meses tanto no horizonte do passado quanto no do futuro, em relação ao ponto de observação, conforme já ilustrado na Figura 4.

Além disso, foram consideradas 9 safras, sendo que cada uma com o ponto de observação em um diferente mês. A intenção inicial era analisar 12 safras, abrangendo todos os meses, contudo não houve histórico suficiente na empresa. Portanto, considerou-se suficiente apenas 9 safras, seja pelo tamanho da amostra e por presumidamente abranger alguma sazonalidade, se assim houvesse. Para a marcação de bom e mau pagador na amostra, considerou-se todo o período de observação, ou seja, uma vez ocorrida a marcação de mau pagador após o ponto de observação, ela permanecerá, independe se o cliente tornou-se bom ainda dentro desse período. Ou seja, uma vez marcado como mau pagador no período de observação, essa informação será gravada e permanecerá.

Dessa forma, apresenta-se na Tabela 3 o total de clientes na amostra, conforme a data do ponto de observação, e a quantidade por tipo de cliente.

Tabela 3 - Número total de clientes na amostra, conforme a data do ponto de observação e a quantidade por tipo de cliente.

	Data da Observação	Total de Clientes da Amostra	Bom	Mau	Indeterminado
Base desenvolvimento	abr/10	7.449	6.541	450	458
	mai/10	8.638	7.664	523	451
	jun/10	9.089	8.185	425	479
	jul/10	10.913	9.893	513	507
	ago/10	10.985	10.009	469	507
	set/10	9.186	8.296	431	459
	out/10	10.258	9.203	509	546
	nov/10	11.827	10.673	545	609
	dez/10	13.521	12.138	670	713
	Total	129.427	82.602	4.535	4.729

Por exemplo, foram 13.521 clientes que tiveram ponto de observação em dezembro de 2010, sendo que as variáveis históricas que predizem seu comportamento na instituição foram coletadas de junho de 2010 até o início de dezembro de 2010. Já o comportamento do cliente será medido nos 6 meses posteriores, ou seja, de janeiro de 2011 até junho de 2011.

Esses são os valores já desconsiderando observações com inconsistência no preenchimento. As observações com problemas totalizaram menos de 1% da base inicial. A partir de então, foram separadas amostras para análise e teste, considerando todos os 4.535 clientes identificados como maus e uma amostra aleatória para selecionar outros 4.535 clientes identificados como bons do total de 82.602 clientes bons na amostra inicial, formando uma amostra com 50% de cada grupo, totalizando 9.070 clientes para o estudo.

Visto que a verdadeira proporção de maus na população em estudo é muito baixa (em torno de 5%), poderia haver problema de discriminação na amostra, dado que a maioria seria de bons clientes. Por este motivo, todos os maus foram utilizados e apenas uma amostra de bons clientes. Contudo, desconsiderou-se a verdadeira proporção entre os grupos para uma correção no resultado final. Na finalização, é realizado um ajustamento com a *priori* da verdadeira proporção, através de um recurso no processo de decisão disponível no *SAS Enterprise Miner*.

Com base na amostra final, de 9.070 clientes, foi separada de forma aleatória, a amostra de análise, utilizada para a construção do modelo; e a amostra de teste, utilizada para o teste do modelo, na proporção de 70% e 30%, respectivamente.

Análise dos dados

Houve necessidade de codificar os valores que estavam marcados como *missings*, quando na verdade indicavam a nulidade da respectiva variável. Oito variáveis cadastrais e três históricas foram excluídas por apresentarem mais de 50% de *missings.* Outra análise importante é dos *outliers*, pois em dados financeiros é comum haver muita discrepância nas variáveis. O tratamento dos *outliers* ocorreu no momento da modelagem via *SAS Enterprise Miner*, através da função *Replacemente,* que facilita a recodificação dos valores discrepantes. Esses valores podem ser considerados *missings*, ou incorporados considerando-o como sendo pertencentes aos quartis extremos da distribuição da variável, ou ainda estabelecer um limite de forma manual. Nessa etapa é muito importante a opinião de um analista para que identifique o que realmente é um *outlier* ou é uma informação incorreta. Obtiveram-se, assim, cinco variáveis que foram limitadas, consideram os valores fora do intervalo como sendo *missings;* foi o caso de tempo do último depósito com valores negativos, tornando os valores negativos *missings,* por exemplo. Para outras 32 variáveis foram recodificados os valores discrepantes, ao exemplo da soma de investimento que foi limitada a R$500.000, pois havia poucos casos que destoavam deste valor.

Análise bivariada

Nessa seção analisa-se a relação entre a variável resposta, que identifica o bom e o mau cliente e as demais variáveis. Nessa fase, vinte e sete variáveis foram rejeitadas da análise por terem poder de discriminação muito próximo de zero, esse poder foi medido pela estatística de Gini e pelo *Information Value*. Porém, seis dessas continuaram por serem importantes, segundo experiência de analistas.

Para incluir algumas variáveis, como CEP residencial, na análise foi necessário categorizá-las, devido ao elevado número de atributos de cada uma delas. Para tal agrupamento foi utilizado o apoio da função *Interactive Grouping* disponível no *SAS Enterprise Miner*, onde se categoriza as variáveis conforme o peso da evidência (*WOE*). Após, cada grupo formado transforma-se em uma variável *dummy* que serão, portanto, as variáveis preditoras para a construção do modelo. Com esse artifício evitam-se problemas decorrentes da não linearidade dos atributos no cálculo da análise multivariada. Com isso obteve-se um total de cento e quatorze variáveis *dummies*, além das treze que já eram binárias, transformadas em *dummies*, e ainda foram mantidas vinte e duas variáveis intervalares, além da variável resposta com a informação de bom e mau cliente, e a variável chave que identifica o cliente, totalizando assim 151 variáveis para o início da análise.

Obtenção da fórmula preliminar

Para a construção do modelo, tanto para a parte da regressão logística quanto para a parte da rede neural, o *software* utilizado foi o *SAS enterprise Miner* versão 6.1 e 6.2. Na construção do modelo de regressão logística, utilizou-se o método *stepwise*, com 0,05 de significância para entrada e saída de variáveis. Como vantagem, esse método proporciona ação corretiva para o problema de multicolinearidade, pois desconsidera variáveis que apresentam sinais de multicolinearidade, optando por manter no modelo as de maior significância (SELAU, 2011). Algumas variáveis foram reagrupadas, a fim de facilitar a entrada no modelo. Não foram adicionados efeitos de interação de variáveis no modelo a fim de melhorar a compreensão do modelo final. Segundo SICSÚ (2010), modelos de *Credit Scoring* encontrados no mercado não têm utilizado essa prática.

Para a composição do modelo logístico, 32 variáveis *dummies* e 5 intervalares foram significativas, sendo que 7 *dummies* e 1 intervalar proviam das informações cadastrais e 25 *dummies* e 4 intervalares são variáveis de comportamento obtidas na observação do cliente durante o período histórico. Pôde-se perceber que as variáveis históricas agregam mais informação que as variáveis cadastrais, ajudando no poder de discriminação do modelo.

Para preservar as informações da empresa em que se está realizando o estudo, as variáveis serão apresentadas na fórmula final de forma codificada, permitindo a identificação se a variável é de origem cadastral ou histórica. O código obedece até quatro dígitos: XYZW, sendo que se X=D a variável é dummy, se X=I a variável é

intervalar; se Y=C a variável é cadastral, se Y=H a variável é histórica; Z é o número identificador da variável e W indica o grupo da variável, quando essa for *dummy*. Supondo que estado civil seja a variável identificada com o número 3 (Z=3), essa variável é proveniente das informações cadastrais (Y=C), além disso, ela é uma variável *dummy* (X=D) e possui dois grupos (W=1 e W=2), logo obtém a decomposição DC31 e DC32. São esses códigos contidos na equação 07, que apresenta a fórmula obtida com a regressão logística.

$$P(Y=1) = \frac{1}{1 + \exp(-1{,}6295 - 0{,}0934\,DC11 + 0{,}8232\,DC21 + 0{,}1951\,DC22 - 0{,}3816\,DC25 + 0{,}2350\,DH32 + 0{,}3193\,DH33 - 0{,}1873\,DC42 + 1{,}2688\,DH51 + 0{,}2577\,DH52 - 0{,}2198\,DH54 - 0{,}6264\,DH55 + 0{,}3493\,DH71 + 0{,}0210\,DH72 - 0{,}0470\,DH81 + 0{,}1791\,DH82 + 0{,}5048\,DH91 + 0{,}1183\,DH92 - 0{,}3361\,DH93 - 0{,}5991\,DH101 - 0{,}3137\,DH102 - 0{,}0703\,DH102 + 0{,}3694\,DH104 - 0{,}8232\,DH111 - 0{,}1534\,DH112 + 0{,}2203\,DH113 + 0{,}2555\,DH114 - 0{,}7681\,IH12 - 0{,}0932\,IH13 - 0{,}0127\,IH14 + 0.0142\,IC15 - 0.1359\,IH16 - 0.1291\,DC17)} \quad (07)$$

As 14 variáveis com sinais positivos são associadas com ser bom pagador e as 17 de sinais negativos com ser mau pagador. Assim é possível identificar o perfil do cliente desejado; se o cliente é casado, possui relacionamento com a empresa por mais de um ano, não apresentou cheque devolvido nos últimos seis meses, possui veículo, entre outras, tem maior probabilidade de ser um bom pagador.

Todas as variáveis e resultados que compuseram o modelo final de regressão logística servirão como nós de entrada da rede neural. Dessa forma, o modelo final será um hibrido da regressão logística com a rede neural. Assim, garante-se a melhora na decisão da estrutura da rede, além de dar um suporte às dificuldades de interpretação dos resultados obtidos.

A rede neural foi construída através da função de ativação sigmóide e o algoritmo de aprendizado supervisionado de retropropagação de erro. Várias redes foram criadas, todas usando a regressão logística como entrada, diferenciando-se nas quantidades de neurônios na camada oculta para verificar o desempenho quanto à predição dos bons e maus clientes. Sendo assim, a RN2 foi a melhor rede, tanto para a amostra de análise, quanto para a de teste, conforme os resultados apresentados na Tabela 4, com as três melhores redes construídas.

Tabela 4 - Comparação dos melhores modelos neurais construídos.

Modelo	Nº Neurônios camada oculta	KS	
		Análise	Validação
RN1	28	61	58
RN2	30	62	61
RN3	35	62	60

Análise da validação da fórmula de pontuação

Depois de concluídas as duas partes do modelo hibrido, regressão logística unida com a rede neural, verifica-se o quão eficiente tornou-se o modelo. Para isso utilizou-se indicadores que irão embasar a conclusão: (i) estatística de Kolmogorov-Smirnov (KS); (ii) área abaixo da curva ROC (AUC); (iii) índice de Gini. Todas as medidas devem ser avaliadas tanto na amostra de análise, utilizada para o desenvolvimento do modelo, como na amostra de teste, necessária para garantir que o modelo seja adequadamente utilizado para previsão. Pela natureza das variáveis, os modelos de *Behavioral Scoring* apresentam melhores resultados em comparação aos modelos de *Aplication Scoring*.

No estudo aqui apresentado, foi proposta uma nova abordagem de modelagem utilizando um modelo híbrido. Dessa forma, para verificar o poder de predição da hibridação das técnicas (regressão logística seguida de rede neural) em relação à regressão logística pura, apresenta-se a Tabela 5 expondo as comparações. Nela percebe-se que houve um acréscimo significativo em todos os indicadores para o modelo híbrido.

Tabela 5 - Comparação dos modelos.

Indicador	Regressão logística	Regressão logística + Rede neural
KS	59	61
ROC	0,87	0,94
GINI	0,73	0,76

Em termos do valor de KS, o modelo híbrido conseguiu alcançar um nível muito bom de diferença entre as distribuições acumuladas de bons e maus clientes, sendo que a regressão logística também apresentou um nível bom. Alguns pontos a mais nessa diferença podem significar aumento nos lucros para uma empresa, portanto essa diferença deve ser considerada e se possível medida em termos de retorno monetário. O valor da área sobre a curva ROC no modelo híbrido também foi bastante expressivo, indicando que a capacidade de identificar corretamente os maus créditos (sensibilidade), assim como a capacidade de identificar os bons créditos (especificidade) está bem ajustada, chegando bem próximo a 1, valor máximo. Também o índice de Gini, que sumariza o desempenho do modelo sobre todos os pontos de corte, apresentou melhora no modelo híbrido.

5. Conclusões

Esse artigo apresentou uma sistemática para a construção de um modelo de *Behavioral Scoring*, propondo um processo de modelagem híbrida de dois estágios com regressão logística e redes neurais. Todos os passos para a obtenção do modelo foram abordados, tanto em relação à obtenção das variáveis e amostras, quanto na abordagem das técnicas. Esse trabalho inova não só no detalhamento do processo de construção de um *Behavioral Scoring,* como também quanto à abordagem de modelos híbridos, que estão sendo recentemente estudados internacionalmente, sendo esse um dos precursores na apresentação dessa técnica com dados brasileiros. Dessa forma, o modelo aqui desenvolvido pode servir de apoio para pesquisadores e analistas de empresas que desejam desenvolver seus modelos.

A técnica de modelagem híbrida aqui desenvolvida foi condizente com estudos já realizados, apresentando superioridade à tradicional (regressão logística). Além disso, com o apoio dos resultados da regressão logística, como nós de entrada da rede neural, técnica que vem sendo cada vez mais utilizada, contornaram-se as características indesejáveis das redes neurais, como processamento lento e dificuldade na interpretação das variáveis. Cabe ressaltar que, a busca por uma rede neural mais eficiente é de suma importância e depende da experiência do pesquisador, visto que algumas redes treinadas (RN1), tendo como nós de entrada os resultados da regressão logística, alcançaram os mesmos indicadores de eficiência da regressão logística, sendo aconselhável, nesses casos o uso dos modelos mais simples que alcançam o mesmo resultado, ou seja, o modelo mais parcimonioso.

Portanto, a utilização de modelos de previsão de risco de crédito que utilizam as variáveis comportamentais dos clientes elimina a subjetividade da análise tradicional, aproveitando as informações ricas do comportamento do cliente que se encontram armazenadas em bancos de dados, muitas vezes inutilizáveis. Além disso, a padronização do procedimento de decisão e a velocidade na análise do crédito são ganhos que aumentam a rentabilidade da empresa, garantindo uma maior eficiência no atendimento dos clientes.

Durante a execução desse trabalho surgiram algumas questões que não foram abordadas nesse artigo e a seguir serão apresentadas como sugestões para trabalhos futuros: (i) estudar o impacto do uso de grupos de clientes bons e maus de tamanhos iguais ou diferentes na estimação e previsão do modelo; (ii) avaliar, a partir

de uma medida de retorno monetário, o quanto o aumento no valor de KS de 59 (regressão logística) para 61 (regressão logística + rede neural) pode trazer de retorno que se justifique o uso de modelos híbridos; (iii) utilizar algoritmos genéticos para encontrar os parâmetros ótimos da redes neurais e a quantidade ótima de neurônios da camada oculta; (iv) aplicar o modelo construído para avaliar os ganhos reais em termos de redução da inadimplência; (v) utilizar métodos *ensemble* (TWALA, 2010; WANG *et al*., 2011) em modelos de *Behavioral Scoring.*

Referências Bibliográficas

ABDOU, H. A. An evaluation of alternative scoring models in private banking. **The Journal of Risk Finance**, v.10 n.1, p.38-53, 2009.

AKHAVEIN, J.; FRAME, W.S; WHITE, L. J. The diffusion of financial innovations: An examination of the adoption of small business credit scoring by large banking organizations. **The Journal of Business**, v.78, n.2, p.577-596, 2005.

BAESENS, B.; SETIONO, R.; MUES, C.; VANTHIENEN, J., Using neural network rule extraction and decision tables for credit-risk evaluation. **Management Science**, v.49, n.3, p.312-329, 2003.

BEE, W. Y.; SENG, H. O.; NOR, M. H. Using data mining to improve assessment of credit worthiness via credit scoring models. **Expert Systems with Applications**, v.38, p.13274-13283, 2011.

BUENO, V. F. F. **Avaliação de risco na concessão de crédito bancário para micros e pequenas empresas**. Florianópolis: UFSC, 2003. Dissertação (Mestrado em Engenharia da Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, 2003.

CHEN, W.; MA, C.; MA, L. Mining the customer credit using hybrid support vector machine technique. **Expert Systems with Applications**, v.36, p.7611-7616, 2009.

CHENG, L. H.; MU, C.C; CHIEH, J.W. Credit scoring with a data mining approach based on support vector machines. **Expert Systems with Applications**, v.33, p.847-856, 2007.

CHUNG, H. M.; GRAY, P. Special section: data mining. **Journal of Management Information Systems**, v.16, n.1, p.11-16, 1999.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada: para cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2007.

DYCHE, J.; DYCH, J. **The CRM handbook: a business guide to customer relationship management**. Reading, MA: Addison-Wesley, 2001.

EMEL, B. A; ORAL, M; REISMAN, A.; YOLALAN, R. A credit scoring approach for the commercial banking sector. ***Socio-Economic Planning Sciences***, v.37, p.103-123, 2003.

FINLAY, S. Credit scoring for profitability objectives. **European Journal of Operational Research**, v.202, 2010.

GANG, W.; JIAN, M.; LIHUA, H.; KAIQUAN, X. Two credit scoring models based on dual strategy ensemble trees. **Knowledge-Based Systems**, v.26, p.61-68, 2011.

GHODSELAHI, A. A hybrid support vector machine ensemble model for credit scoring. **International Journal of Computer Applications**, v.17, n.5, p. 975-8887, 2011.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados**. 5.ed. Porto Alegre: Bookman, 2005.

HAND, D.J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of the Royal Statistical Society. Series A (Statistical in society)**, v.160, p.523-541, 1997.

HAYKIN, S. **Redes neurais: princípios e prática**. Trad. Paulo Martins Engel. 2.ed. Porto Alegre: Bookman, 2001.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

HSIEH, N. C. An integrated data mining and behavioral scoring model for analyzing bank customers. **Expert Systems with Applications**, v.27, p.623-633, 2004.

HSIEH, N.C., Hybrid mining approach in the design of credit scoring models. **Expert Systems with Applications**, v.28, p.655-665, 2005

HUANG, Z.; CHEN, H.; HSU, J. C.; CHEN, H. W.; WU, S. Credit rating analysis with support vector machines and neural networks: a market comparative study. **Decision Support Systems**, v.37, p.543-558, 2004

JORION, P. **Value at risk: the new benchmark for controlling market risk**. New York, Mc Graw Hill, 1997.

KIM, M.K.; SOHN, S.Y. Cluster-based dynamic scoring model. **Expert Systems with Applications**, v.32, p.427-431, 2007.

LAHSASMA, A.; AINON, N. R.; WAH, Y. T. Credit scoring models using soft computing methods: a survey. **The International Arab Journal of Information Technology**, v.7, n.2, p.115-123, 2010.

LEE, T.; CHIU, C.; LU, C.; CHEN, I. Credit scoring using the hybrid neural discriminant technique. **Expert Systems with Applications**, v.23, n.3, p.245-254, 2002.

LEE, T.S.; CHEN, I. F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. **Expert Systems with Applications**, v.28, p.743-752, 2005.

LEWIS, E. M. **An introduction to credit scoring**. San Rafael: Fair, Isaac and Co., Inc. 1992.

LOESCH, C.; SARI, S.T. **Redes neurais artificiais: fundamentos e modelos**. Blumenau, FURB, 1996.

LOUZADA, F.; AMARAL, G. J. A.; GUIRADO, L.; SILVA, P. H. F.; ABREU, H. J. ; FERREIRA, M. R. P. Medidas estatísticas da capacidade preditiva de modelos de classificação em credit scoring. **P@rtes** (São Paulo), v. 68, p.7-28, 2008.

MCCLELLAND, J. L.; RUMELHART, D. E. & the PDP research group. **Parallel distributed processing: explorations in the microstructure of cognition**. v. II. Cambridge, MA: MIT Press, 1986.

MCNAB, H.; WYNN, A. **Principles and practice of consumer credit risk management**, CIB Publishing, Canterbury, 2000.

MENDES FILHO, E. F.; CARVALHO, A. C. P. L. F.; MATIAS, A. B. Utilização de redes neurais artificiais na análise de risco de crédito a pessoas físicas. **In: III Simpósio Brasileiro de Redes Neurais**, Recife. Anais. 1996.

MORRISON, A. D, The economics of capital regulation in financial conglomerates**. Geneva Papers on Risk and Insurance: Issues and Practice**, 2003.

SARLIJA, N.; BENSIC, M.; SUSAC, Z. M. Comparison procedure of predicting the time to default in behavioural scoring. **Expert Systems with Applications**, v.36, p.8778-8788, 2009.

SAUNDERS, A. **Medindo o risco de crédito: novas abordagens para value at risk e outros paradigmas.** Rio de Janeiro: Qualitymark, 2000.

SECURATO, J. R. **Crédito: análise e avaliação de risco**. São Paulo: Saint Paul, 2002.

SELAU, L. P. R. **Construção de modelos de previsão de risco de crédito.** Porto Alegre: UFRGS, 2008. Dissertação (Mestrado em Engenharia da Produção), Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, 2008.

SELAU, L. P. R.; RIBEIRO, J. L. D. Systematic approach to construct credit risk forecast models. **Pesquisa Operacional**, v.31, p.1-17, 2011.

SICSÚ, A. L. Credit scoring: desenvolvimento de um sistema de credit scoring – Parte II. **Tecnologia de Crédito**. São Paulo: Serasa, n.5, 1998.

SICSÚ, A. L. **Desenvolvimento, implantação, acompanhamento**. São Paulo: Blucher, 2010.

STEINER, M. T. A.; CARNIERI, C.; KOPITTKE, B. H.; STEINER NETO, P. J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. **Revista de Administração**, v.34, n.3, p.56-67, 1999.

SUSTERSIC, M.; MRAMOR, D.; ZUPAN, J. Consumer credit scoring models with limited data. **Expert Systems with Applications**, v.36, p.4736-4744, 2009.

THOMAS, L.C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. **International Journal of Forecasting**, v.16, p.149-172, 2000.

TSAI, C.F.; CHEN, M.L. Credit rating by hybrid machine learning techniques. **Applied Soft Computing**, v.10, p.374-380, 2010.

TWALA, B. Multiple classifier application to credit risk assessment. **Expert Systems with Applications**, v.37, p.3326-3336, 2010.

WANG, G.; HAO, J.; MA, J.; JIANG, H. A comparative assessment of ensemble learning for credit scoring. **Expert systems with applications**, v.38, p.223-230, 2011.

WEST, D. Neural network credit scoring models. **Computers and Operations Research**, v.27, n.11-12, p.1131-1152, 2000.

WONG, B. K.; SELVI, Y. Neural network applications in finance: a review and analysis of literature. **Information & Manegement**, v.34, p.129-129, 1998.

YU, X.; EFE, M. O.; KAYNAK, O. A general backpropagation algorithm for feedforward neural networks learning. **IEEE Trans**. **on Neural Networks**, v.13, n.1, p.251-254. 2002.

AN ALTERNATIVE APPROACH FOR SCORING BEHAVIORAL USING HYBRID MODELING TWO-STAGE WITH LOGISTIC REGRESSION AND NEURAL NETWORKS.

Abstract

With the progressive growth in lending volumes in Brazil, companies are seeking improvement in assertiveness and flexibility in granting credit analysis, not only for

new customers but also to existing clients. Multivariate techniques have diffused to build predictive models of credit risk that, based on both registration information, and also in the history of the customer relationship in the company, predicting a pattern of risk behavior. The aim of this paper is to propose a system for building predictive models of credit risk based on behavioral data (Behavioral Scoring), using a process of two-stage hybrid modeling with logistic regression and neural networks and evaluate their performance. All stages of construction of the model are discussed in detail, being approached from planning and definition of the model to the analysis of the validation of the scoring formula. The model was applied to a sample of 9070 customers of a financial institution of national performance. The results for this particular study showed that the developed hybrid modeling technique was superior to traditional, stressing that the support of the logistic regression results as input nodes of neural network bypassed the undesirable characteristics of neural networks, such as slow processing and difficulty in interpretation of the variables.

Keywords: credit scoring, behavioral scoring, hybrid modeling, logistic regression, neural networks.